# Construction of a Phonotactic Dialect Corpus using Semiautomatic Annotation*

*Reva Schwartz*[1], *Wade Shen*[2], *Joseph Campbell*[2], *Shelley Paget*[3], *Julie Vonwiller*[3], *Dominique Estival*[3], *and Christopher Cieri*[4]

[1] United States Secret Service, Washington, DC USA
[2] MIT/Lincoln Laboratory, Lexington, MA USA
[3] Appen Pty Limited. Sydney NSW Australia
[4] Linguistic Data Consortium, Philadelphia, PA USA

Reva.Schwartz@usss.dhs.gov, {swade,jpc}@ll.mit.edu,
{spaget,jvonwiller,destival}@appen.com.au, ccieri@ldc.upenn.edu

## Abstract

In this paper, we discuss rapid, semiautomatic annotation techniques of detailed phonological phenomena for large corpora. We describe the use of these techniques for the development of a corpus of American English dialects. The resulting annotations and corpora will support both large-scale linguistic dialect analysis and automatic dialect identification. We delineate the semiautomatic annotation process that we are currently employing and, a set of experiments we ran to validate this process. From these experiments, we learned that the use of ASR techniques could significantly increase the throughput and consistency of human annotators.

**Index Terms**: corpora, annotation, phonology, dialect identification, linguistic variation

## 1. Introduction

In the field of forensic phonetics, applied phoneticians routinely identify speakers from phonetic and phonotactic characteristics that are hypothesized to be speaker specific. Quantifiable norms of language- and dialect-dependent features are necessary for forensic examiners to assess if a given phonological or phonetic feature is speaker specific or commonly found in that speaker's dialect.

Historically, the sociolinguistic study of dialect variation in speech has been limited to small-scale population studies, usually conducted through sociolinguistic interviews [1]. Dialect analysis of large populations has generally been difficult because the cost associated with collection, transcription, and annotation (what sociolinguists call coding) of speech data has been prohibitive. Counter-examples are few but include the Atlas of North American English [2] and one of the author's [3] previous works on annotation of large-scale corpora of conversational telephone speech.

In this paper, we describe the protocols we are using to collect and annotate a new corpus of phonological variation across dialects of American English. Our goal is to eventually build a large annotated corpus sufficient for establishing dialect norms for a variety of phonological phenomena. We believe that such a corpus will help forensic speech scientists and sociolinguists to quantify variation between and within American English dialects. Furthermore, we expect that this corpus will support research in methods for automatic dialect identification from speech. In sections 3 and 4, we describe the goals of this annotation effort and these protocols in detail.

As annotation is the largest hurdle in terms of time and cost, we employ a semiautomatic process using automatic speech recognition (ASR) techniques. In this paper, we compare the efficiency of this approach with a detailed phone transcription effort and we show that the use of this type of process increases transcriber throughput. It does so by reducing the amount of audio to be annotated by 90%. Additionally, we describe a pilot study comparing two different human annotation protocols, phonological annotation, and their respective consistency rates.

## 2. Corpus Description

To reduce cost and schedule, we sought existing data representative of variability found in American English dialects. Because our methods (described below) require audio recordings, word-level transcripts (not necessarily aligned), dialect labeling of utterances and, foremost, statistically significant amounts of scientifically controlled data (for both measurement and estimation of norms of various linguistic features and development of automatic dialect identification systems), we limited our initial effort to two well-studied dialects of American English; namely, African-American Vernacular English (AAVE) [4] and non-AAVE. This oversimplifies the dialect landscape, but can prove our method and allow rich postanalysis via metadata.

Materials from conversational speech (not necessarily between familiar talkers) and relatively free of accommodation effects were preferred. We sought data from many talkers having multiple recording sessions over time so as to include intra- and interspeaker variability. This allows for the observation of salient features that are rare or infrequent. We chose talkers involved in natural conversation with other talkers of the same dialect to minimize dialectal accommodation. For ease of processing, telephone-quality recordings of separable conversation sides were preferred.

We acquired materials from talk radio and television, television shows and movies with closed captioning, sociolinguistic interviews, and data from publicly available corpora. The bulk of our corpus is derived from the Mixer Corpora produced by the University of Pennsylvania's Linguistic Data Consortium [5] because it best met the aforementioned needs. The Mixer corpus of conversational telephone speech is huge and, although designed primarily for automatic speaker recognition research, includes metadata of

*This work is sponsored by the Department of Homeland Security under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

# Report Documentation Page

*Form Approved*
*OMB No. 0704-0188*



| 1. REPORT DATE | 2. REPORT TYPE | 3. DATES COVERED |
|---|---|---|
| **2007** | **N/A** | **-** |

| 4. TITLE AND SUBTITLE | 5a. CONTRACT NUMBER |
|---|---|
| **Construction of a Phonotactic Dialect Corpus using Semiautomatic Annotation** | 5b. GRANT NUMBER |
| | 5c. PROGRAM ELEMENT NUMBER |
| 6. AUTHOR(S) | 5d. PROJECT NUMBER |
| | 5e. TASK NUMBER |
| | 5f. WORK UNIT NUMBER |
| 7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **MIT Lincoln Laboratory, Lexington, MA, USA** | 8. PERFORMING ORGANIZATION REPORT NUMBER |
| 9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) | 10. SPONSOR/MONITOR'S ACRONYM(S) |
| | 11. SPONSOR/MONITOR'S REPORT NUMBER(S) |

12. DISTRIBUTION/AVAILABILITY STATEMENT
**Approved for public release, distribution unlimited**

13. SUPPLEMENTARY NOTES
**The original document contains color images.**

14. ABSTRACT

15. SUBJECT TERMS

| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT | 18. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON |
|---|---|---|---|---|---|
| a. REPORT | b. ABSTRACT | c. THIS PAGE | | | |
| **unclassified** | **unclassified** | **unclassified** | **UU** | **4** | |

**Standard Form 298 (Rev. 8-98)**
Prescribed by ANSI Std Z39-18

the talker's self-reported race, sex, age, occupation, education, languages, and places where born and raised. Mixer, however, has some limitations: it is not fully transcribed, its talkers are unfamiliar with each other (this has pros and cons), and its dialect classification information is gleaned from self-reported information. Disadvantages aside, Mixer is the single largest dialect triaged corpus that we know of. As such, it contains the large amounts of high-quality telephone speech exhibiting both inter- and intraspeaker variability. Furthermore, Mixer collections are ongoing and it might be possible to collect additional dialects, languages, and demographic information in the future.

For our pilot study, a small portion of Mixer was selected, consisting of five 5-minute telephone conversations (10 conversation sides; much larger portions are available for expanded work). Three conversations (6 sides) were used to acclimate the trained human annotators to the dialect and develop the tools and work flow. The two remaining conversations (4 sides), with manual transcription and alignment, were used in the pilot-study annotation work described in the following sections.

In addition to Mixer, through contacts within the sociolinguistic research community, we have gathered 7.41 hours from 13 speakers of annotated AAVE corpora from Louisiana [6] and 5 hours of North Carolina annotated AAVE data [7]. Additionally, 9 hours of annotated television data have been collected. These data sets might also undergo the phonetic annotation processes described below.

## 3. Annotation Goals

For the speech data in this corpus, we would like to produce annotations of phonotactic phenomena that potentially differ across dialects. This can be accomplished in a number of ways: 1) detailed phonetic transcription of all speech or 2) phonetic transcription of divergences in dialect data. In this paper, we attempt to design an annotation protocol that examines only potential dialect divergences. In this way, it is possible to minimize the cost of annotation and increase the amount of relevant data that we can annotate. We start with data that has been preclassified by dialect and, possibly, pretranscribed at the word level. Our goals for this protocol are described in detail as follows.

**Low cost annotation:** In this pilot project, we sought highly detailed annotations of speech, with all attention specifically paid to the phonological level. In working with annotators at Appen and LDC, we discovered that it is possible to automate some of the process and, therefore, increase efficiency. The trade-off is between high accuracy and high speed. For the most accurate transcriptions, human annotators should phonetically transcribe speech. As a lower cost alternative, we would like to use word-level transcripts segmented at the phrase level – usually between breath groups. From this point, ASR technology can be used to extract possibly errorful phone-level alignments, thereby reducing the complexity of the annotation task. As this process is imperfect, it is still necessary to correct the output from ASR systems. This process is described in more detail in Section 4.

**Identification of dialect-specific phonetic/phonological effects:** This is, of course, the primary goal: detailed labeling of variations from General American English (GAE). Dialect differences can occur across different linguistic levels, not just at the phonotactic and phonological levels, but also in the lexicon, morphology, syntax, etc. Conversational speech, transcribed at the word level, does not always use standard orthographic form. Some transcriptionists attempt to represent pronunciation differences within orthographic transcripts (e.g., "aks" for "ask"), while others use the standard orthography in all cases and still others permit a small, controlled set of conventionalized variants (e.g., "gonna" in place of "going to"). Most of the transcripts we have encountered so far subscribe to the third practice. Therefore, we assume in this effort that transcripts use the standard orthography with a few exceptions, but that variations in lexical choice, morphology and syntax are transcribed as they are actually spoken (including disfluencies), not as what was expected. This is generally true of speech data transcribed for ASR. Variations that occur at or above the lexical level are assumed to be preannotated. Our remaining challenge is to locate dialect-specific phonetic/phonological effects. For that, we need annotators to identify variations from the forms found in a GAE pronunciation lexicon. Decisions were made to update the pronunciation dictionary very rarely and only in cases when transformations seem to occur on a somewhat frequent basis in GAE.

**Annotation consistency:** The resulting annotations are only useful if a high level of consistency can be obtained across annotators. Prior to any annotation process, all annotators must agree to a set of transcription protocols, with special instructions for items such as transformations, pauses, incomplete utterances, etc. To train on these protocols, we asked that all annotators take part in a small test for intra- and interannotator consistency. This would entail the annotators transcribing the same, short sample (we used approximately 2 minutes) of speech using the agreed-upon protocols. Post-test evaluation might uncover inconsistencies and places for improvement.

## 4. Annotation Process

Our process is designed to minimize the amount of data each human annotator needs to annotate. To this end, we employ a series of automatic preprocessing steps. Beginning with human transcribed audio files, we performed three stages of preprocessing using a standard HMM-based ASR system. First, files were segmented to identify speech segments through a process of forced alignment using ASR. Using the transcript corresponding to each file, we allow silence to be inserted between each word. Segments were then extracted by breaking the transcript when silences with length greater than 0.3 seconds were detected (approximate breath groups).

Using this segmentation, we realigned the audio to the transcript. During this pass, we produce phone-level alignments that are then augmented with syllable and word boundary information. Utterances with sufficiently low average likelihood ASR scores (per frame) are rejected and not considered for further processing and annotation.

As it would be costly to have human annotators examine the alignment of every phone, we automatically extract regions for which we expect phonetic transformation. These *regions of interest (ROIs)* are marked using an automatic tool that, given a set of phonetic/phonological transforms per dialect and the true dialect of the audio file being annotated, finds regions in which a phonetic/phonological transform could have applied. *Table 1* shows a number of these rules (with TIMIT phones) and example ROIs that each rule could induce. For our pilot study, we created a superset of rules

derived from phenomena described in the sociolinguistic literature [8][9][10][11].

| Rule | Example ROI |
|---|---|
| [r]->0 / [+vowel] ___ [+vowel] | p ax r ae m ax s |
| 0->[ax] / [k,t,s] ___ [z] # | B ae k z # |

*Table 1: Sample Phonetic Transformation Rules*

For each transformation rule, the right side represents the Standard English allophone that is transformed to the dialect-specific form on the left side. It is important to note that this formalism relies upon the familiarity of GAE, using it as a point of departure to describe dialect variations. It is not the authors' intent to assert that all dialect forms derive either diachronically or synchronically from GAE forms. Because these rules are critical for filtering the amount of data each annotator sees, any rule missing from this stage could cause data to be missed. To avoid this, our rules were constructed in three stages: 1) consultation with dialect experts (sociolinguists) and the literature on a per-dialect basis, 2) complete test annotation of a small set of representative data, and 3) addition of new rules as needed.

The resulting ROIs are labeled with the applicable rule(s). Since the application of multiple rules could create overlapping contexts, rules are grouped by types of general phenomena. These are presented to annotators as different tiers. Currently, the set of general phenomena include reduction, epenthesis, substitution, metathesis, rhoticity and /L/-vocalization.

### 4.1. Human Annotation

ROI-marked phone and word transcripts were combined to create a multitiered annotation display for human annotators. Annotators were asked to follow the procedure shown below (simplified):

*For each segment*
Mark voice quality characteristics using agreed upon guidelines.

*For each word*
Choose correct pronunciation variant for each word. If none exists, add the correct variant to the transcript using the phones listed below.

*For each region of interest*
Determine whether the feature label for this region is present and to what extent. The extent to which a feature label is present should be indicated on a scale from 0 to 2 (2 being fully present and 0 being absent).

Annotators were presented annotation tiers using WaveSurfer v. 1.8.5 [12] as shown in Figure 1.

Annotators approached the data focusing on ROIs, looking at one phenomenon at a time. For example, if the annotators were focusing on rhoticity, they would listen to a section of data with predicted rhoticity to ascertain if there were any instances of the expected phenomena; e.g., r-lessness. Annotators would then look in more detail at the point where, for instance, a rhotic was deleted and adjust the word tier, if incorrect. Using a GAE phone set, annotators would also adjust the phone tier to represent the sound that was actually produced by the speaker. For example, if "car" was pronounced as /k aa/ but the automated phone output was /k aa r/, the annotators would remove the /r/ phone, ensure the word and phone tier labels were correctly aligned for that region, and mark in the rhoticity tier (.rho shown in figure 1) the rule that applied. In this case, the rule would be

```
[r] -> 0 / [+vowel] __
```

Approaching the data one phenomena at a time improved consistency and was more time efficient than working through the entire utterance in detail. For further consistency checks and to compare like phenomena, annotated occurrences of the same phenomena were extracted and collated into a separate file to view the degree of difference, for checking internal consistency of the individual annotator and to pick up on acoustic cues. This proved to be especially useful for rhoticity and /L/-vocalization.

Determining the extent to which a rule could be applied

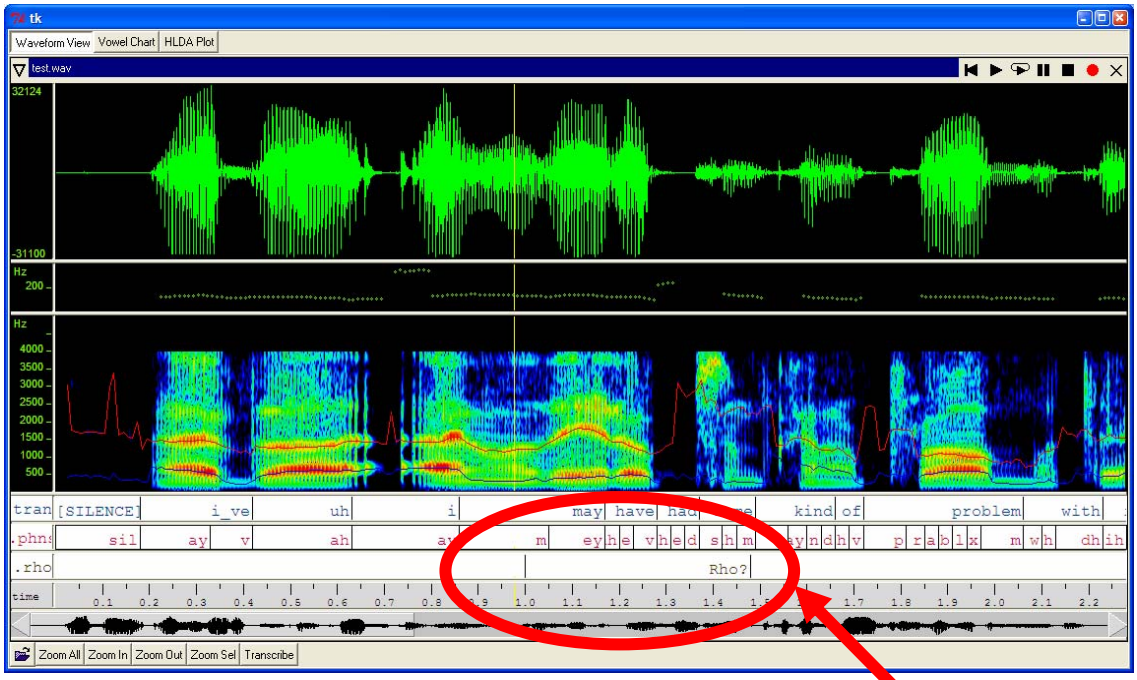


*Figure 1: Example of Tiered Transcriptions*

required the development of a 3-point scale system to specify 1) definite evidence of feature, 2) some evidence of feature or 3) no evidence of feature. For all features, there was a degree of subjectivity involved in determining the scale rating, but in conjunction with spectrographic evidence such as formant structure, it was possible to take a well-balanced approach. In addition to marking up the data for reduction, epenthesis, substitution, metathesis, rhoticity and /L/-vocalization, annotators noted other dialect-specific features that could be used as dialect identifiers, such as intonation and pausing patterns and other conversational techniques.

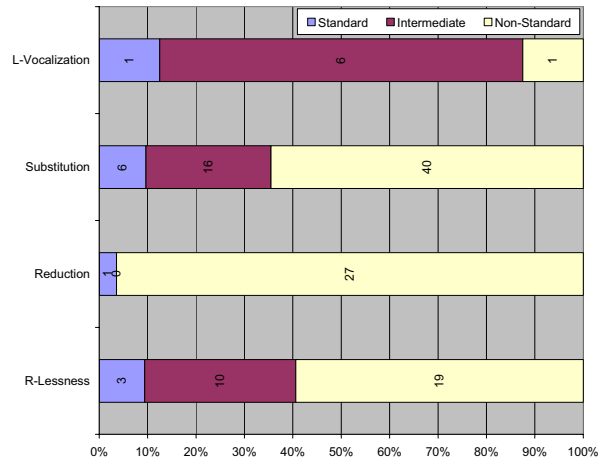


*Figure 2: Distribution of various phenomena from the Pilot Study*

To improve consistency and address some display issues with WaveSurfer, other annotation tools will be investigated that will allow data to be viewed on a per-phenomena basis, rather than from a time-synchronous perspective.

## 5. Pilot Experiments

In the pilot project, one coder manually annotated two 10-minute conversation sides of AAVE data with a restricted set of rules. After speech activity detection, 3 minutes of remaining speech were annotated in detail

In these segments, reduction, substitution, rhoticity and /L/-vocalization proved to be the most productive phenomena. While no instances of epenthesis or metathesis were found, there were contexts that could have been candidates for both phenomena. As this sample was small, more data will be necessary to properly evaluate the frequency and distribution of different phenomena.

While annotating the pilot data, certain frequent features were found that could not adequately be described by the available rules and additional rules were proposed; e.g., nasal plus stop final cluster reduction; word-initial interdental substitution of stop for fricative. Some rules were altered to be more general.

In the 10 minutes of data, a number of the proposed rules were rarely or never observed. For example, consonant cluster reduction occurred infrequently and its distribution was very speaker dependent. Reduction actually occurred in only 13% of potential reduction contexts. Interestingly, one speaker exhibited no instances of cluster reduction (the other speaker reduced 32% of the time). Some initial results for one speaker are presented in the figure 2. For each variable examined, we show the percentage of tokens in which a standard form, dialectal and some intermediate forms were exhibited as a percentage of total ROIs.

Results from this experiment have helped to inform and revise the process for large corpora. We have found that careful annotation of a small subset is useful prior to large-scale annotation. This process allows for revision of the rules needed to find regions of interest and to calibrate the annotation scales on a per-dialect basis.

## 6. Ongoing and Future Efforts

For the pilot phase, preannotated regions of interest were not available, so ROIs were manually identified (to simulate follow-on work) and then annotated at the word, phone and rule levels. The full project will evaluate whether preannotated data will speed up the process. We will address this plus annotation consistency, rules, and the availability of our data in future publications.

There are other potential gains from automation. We expect that using a sampling procedure per dialect and repeating the pilot exercise as described above, we can limit the number of missed phenomena during semi-automatic annotation. If other relevant dialect features emerge during sampling, we can adjust our processing pipeline accordingly.

Future work is planned to incorporate additional automation, features, and rules to study statistically significantly large amounts of data to support large-scale dialect and typicality analysis and automatic dialect identification.

## 7. Acknowledgements

The authors acknowledge the assistance of Malcah Yaeger-Dror and Bridget Anderson for their consultation and data.

## 8. References

[1] Labov, W. 1972. “Some Principles of Linguistic Methodology,” *Language in Society,* p 97-120.

[2] Labov, W., Ash, S. & Boberg, C. 2006. The Atlas of North American English: Phonetics, Phonology and Sound Change. Berlin: Mouton de Gruyter.

[3] Cieri, C. and Strassel, S. 2003. Robust Sociolinguistic Methodology: Tools, Data and Best Practices, a workshop presented at NWAVE 32, Philadelphia.

[4] Wolfram, W. & Thomas, E. 2002. *The Development of African American English (Language in Society).* Blackwell Publishing Professional.

[5] Cieri, C., Andrews, W., Campbell, J., et al. 2006. “The Mixer and Transcript Reading Corpora: Resources for Multilingual, Crosschannel Speaker Recognition Research,” *International Conference on Language Resources and Evaluation (LREC),* ELRA, Genoa, Italy, May 22-28, pp. 117-120.

[6] Dubois, S. & Horvath, B. 2003. “Creoles and Cajuns: A portrait in Black and White.” *American Speech* 78:2, 192-207.

[7] Thomas, E. 2002-04. Corpus of recordings for “Socio-Phonetic Cues Differentiating African American and European American Voices,” National Science Foundation Grant BCS-0213941.

[8] Davis, L. 1983. English Dialectology: An Introduction. Alabama: University of Alabama Press.

[9] Olive, J., Greenwood, A. & Coleman, J. 1993. Acoustics of American English Speech, New York: Springer-Verlag.

[10] Alim, H. 2004. You Know My Steez: An Ethnographic and Sociolinguistic Study of Styleshifting in a Black American Speech Community. Duke University Press.

[11] Wolfram, W. & Schilling-Estes, N. 2005. *American English: Dialects and Variation, 2e. Appendix: An Inventory of Socially Diagnostic Structures.* Blackwell Publishing Professional.

[12] Sjlander K. & Beskow J. 2000. “WaveSurfer - an open source speech tool.” *Proc of ICSLP*, Beijing, Oct 16-20, 4:464-467.